***Can all foods be classified as a salad, soup, or sandwich?***

*A study in classification theory*

Contents

Abstract ……………………………………………………………………………………. 3

Introduction ………………………………………………………………………………... 3

Research review …………………………………………………………………………… 5

Dissertation ………………………………………………………………………………... 10

Evaluation …………………………………………………………………………………. 20

Bibliography ………………………………………………………………………………. 22

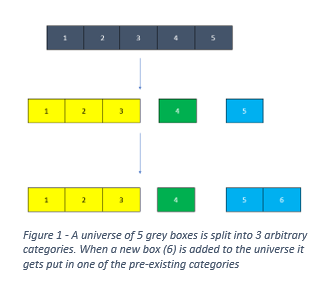
# **ABSTRACT**

The main body of work that goes into classification and categorisation is defining the model that gives the rules of organising and determines which classes elements get placed into. The main goal of this paper is to present a new method of modular model-based classification, building on the framework and ideas set out in *“Ordination on the Basis of Fuzzy Set Theory”* specifically in the domain of ***complete universe categorisation,*** heavily inspired by the work in *“Cognition and Categorization”* as well as influenced by “*Learning visual biases from human imagination”*. Additionally, the author defines the class structure as a concept before going on to analyse individual segments of the question and the properties of the model. Finally, the author employs a method of model analysis formulated using the normal equation for linear regression as a basis. Ultimately the author describes the process of model construction and its applications across categorisation theory using the MSPACE method.

# **INTRODUCTION**

Given an appropriate model, anything can be categorised into any arbitrary group. A common misconception in the field of classification theory is that categorising objects is about choosing the correct labels. Whilst this can be important, the main body of work that goes into classification and categorisation is defining the model that gives the rules of organising and determines which classes elements get placed into. A model can be interpreted as a child’s toy, where different 3D shapes can be pushed through corresponding holes: The actual shapes used (or in our case, the elements being sorted) are irrelevant, but the concept of having an entity that contains not only all the different classes - but their restraints as well, acts as a good analogy for a model. Models contain all the class names, the rules that determine which objects get put into what classes and the variables/properties that every item holds[[1]](#footnote-2)

The main goal of this paper is to present a new method of classification, building on the framework and ideas set out in *“Ordination on the Basis of Fuzzy Set Theory”* [*(****Roberts David, 1986***)](#_Roberts,_D.W._(1986).) specifically in the domain of ***complete universe categorisation*** (*figure 1*). The author also describes a method of structural similarity not seen previously, heavily inspired by the work in *“Cognition and Categorization”* [***(Rosch et al, 1980)***](#_Culbert,_S.S.,_Rosch,) as well as influenced by “*Learning visual biases from human imagination”* [***(Oliva et al, 2015)***](#_Vondrick,_C.,_Pirsiavash,).

Additionally, the author defines the other measures of similarity employed as well as the class structure as a concept before going on to analyse individual segments of the question and the properties of the model.

Finally, the author employs a method of model analysis formulated using the normal equation for linear regression as a basis [***(Uyanık & Güler, 2013)***](#_Uyanık,_G.K._and), as well as discussing alternative methods for reflecting on the fit of your model before going on to describe the newly devised MSPACE system for a step by step approach towards categorisation.

**Research review**

Over the course of my project, I have looked at sources and academic papers in a broad range of topic areas to reinforce and solidify my understanding of my topic. This research review attempts to analyse and evaluate those texts to assess their meaning, impact and validity in the field of classification and categorisation. To aid this examination, I have split my sources into 6 distinct topic groups and will be evaluating each, one at a time. These topic headings are as follows: [**Logic**](#_Logic), [**Set theory**](#_Set_theory), [**Group theory**](#_Group_theory), [**Information systems**](#_Information_systems), [**Classification theory**](#_Classification_theory) and [**Miscellaneous**](#_Miscellaneous) (for sources that didn’t fit into any of the above)

# ***Logic***

The smallest body of research I did was looking at concepts in mathematical logic. I looked at this area because I wanted to gain a better understanding of mathematical models as well as how models can be used to simulate complex concepts in the real world. This was important for me to learn about because in my dissertation I chose to use maths heavily in an area in which it has rarely been employed in the past. The few crossover papers I found (for example:[***Roberts David, 1986***),](#_Roberts,_D.W._(1986).) fall more neatly into another category so I will not discuss them in this section.

In “*The Inescapability of Gettier problems”* ([***Zagzebski Linda, 1994***](#_Zagzebski,_L._(1994).)), Zagzebski describes the problem of perceived truths (either through evidence or extrapolation) being false: a problem that affects models of all kinds, but especially those in classification theory. This was helpful for me in understanding some of the limitations that my model would have. As models are, by tautology, designed to depict the real world, this source was useful to me despite being written by a philosopher: looking at life in general as opposed to mathematics specifically.

The ideas Zagzebski introduces in this paper are also raised in a more general way in [***Philip Welch and Benedikt Löwe’s 2001 paper***](#_Löwe,_B._and)***:*** “*Set-Theoretic Absoluteness and the Revision Theory of Truth”*. Whilst most of this paper was either too advanced mathematically - or irrelevant to my topic – it does bring up some interesting points about the nature of proof. Welch and Löwe argue that it is always easier to disprove something than it is to cover all cases and prove it is true. In the field of classification theory, having a model that accurately reflects the real world is one of the fundamental requirements of any classification (if the model is incorrect then any analysis done on data produced by the system will be inherently wrong). Both papers highlight some important truths about all models: It is an incredibly difficult task to determine whether the model is reflective of life or whether it is just true for the data that it is tested on. Additionally, it is far easier to find a sample unit for which the model breaks down than it is to verify that the model holds true *Omnis*. Succinctly, it is hard to tell if the model works or just “looks” like it works.

In addition to these two main sources, I also explored other papers including “*Mathematical Logic as based on the theory of types”*[*(****Russel Bertrand, 1908****)*](#_Russel,_B._(1908)._1)*.* However, I found this paper was too mathematically advanced for me to fully grasp and it didn’t give me as much insight into the topic as I would have liked.

I found that the paper “*Knowledge and certainty*” *(*[***Stanley Jason, 2008***](#_Stanley,_J._(2008).)*),* was perfect for summing up my findings in the area of mathematical logic. Stanley describes how knowledge needs 3 things to be considered true (or, in my case, how a model requires 3 things to be properly representative): Truth, belief and certainty. The first of these is self-explanatory – for the results of a model to be considered a statement of knowledge, they must have the appearance of recognising integral (and observable) aspects of real life. This means that the data in your model must be applicable in the most basic of situations, otherwise it won’t be the right fit for more complex cases.

The second requirement is harder to describe in the case of mathematical models (a drawback of this paper was its conceptual framing). In this instance, I am interpreting it to be the same as holding a hypothesis. This is the idea that once you have a model that is true for some arbitrary data points, then there must be the belief that that model holds for all items in the universe (since in many situations it is often impractical or impossible to check every unit).

Finally, Stanley argues that an element of certainty is necessary for solidifying findings as facts. In the case of me and my model, certainty is determined as the confidence in the hypothesis created in the above paragraph. It is symbolic of to what extent we are confident in our belief that the model is true for all elements. It is here that problems from the earlier discussions of Zagzebski and Welch & Löwe ’s papers: Even if we have a model and a hypothesis that it holds for everything, we can never reach full certainty that this model doesn’t suffer from Zagzebski’s “Gettier problems” ([***Zagzebski Linda, 1994***](#_Zagzebski,_L._(1994).)), and that a complete demonstration - which would eliminate this possibility – probably cannot be achieved ([***Löwe and Welch, 2001***](#_Löwe,_B._and))

# ***Set theory***

I looked at key concepts of set theory to reinforce my pre-existing knowledge in this area and to look at examples of one of the most basic class structures. I was particularly interested in “fuzzy” set theory – a variant of the standard structure where the degree of class membership is represented by a number – and decided to implement it in my method.

Other than some basic introductory level texts that I looked at to consolidate my understanding of this area, the main paper that really aided my research (and influenced my dissertation) was *“Ordination on the basis of fuzzy set theory”* [(***Roberts David, 1986***)](#_Roberts,_D.W._(1986).). I will touch upon this paper again in the “Classification theory” section of this document, but the main scope of this paper will be discussed here. This paper was especially helpful to me, as it was written specifically for the purposes of classification theory and was one of the few sources that I found which used set theory to accomplish this goal. The beginning of the paper serves as an introduction - or a reminder – to the world of set theory and the nuances with fuzzy sets. Later in the text, Roberts uses this foundation to define a simplistic but efficient system of classification of which my project is a subtype. Most of the core concepts employed in my dissertation are techniques that Roberts used: adapted to be simpler, more adaptable and thus by extension, more generalised. Roberts also produced a follow up paper clarifying some points of confusion and extending his original text [***(Roberts David, 1987)***](#_Roberts,_D.W._(1987).)which I will analyse in another section.

I chose sets as the main data structure for my project partially because it is an area that I was already well acquainted with, but also because it provided me a flexible structure that is easily adaptable and intuitive for all people to understand.

Set theory is a key part of my project but discussing it fully here is beyond the breadth of this research review. Despite being a relatively new area of mathematics, there are lots of great resources to aid in learning the fundamentals. I will go over the core concepts of fuzzy set theory briefly in my dissertation but this itself requires at least a GCSE-equivalent understanding of the basics so some pre-reading may be required (Just the first page of “<https://www.bbc.co.uk/bitesize/guides/z8nfrdm/revision/1>” would be entirely sufficient)

# ***Group theory***

I explored this area as a more complex, higher-level structure than the basic set. I also needed to build up my knowledge in this area to help me better understand and implement some of the information systems that I looked at. Initially, I wanted groups to be the main abstract class structure that I used in my dissertation: However, as I proceeded with my project, I decided that groups weren’t the best fit and I would rarely use them. This was because the nature of groups doesn’t combine well with edge/boundary cases and membership to a group is definite – unlike the fuzzy sets I also looked at. Ultimately, I decided that these factors were too significant for me to rely upon group theory as an integral area of my dissertation and I gradually stopped looking at it.

Most of the texts that I looked at – for example, *“Easy Group Theory”* [***(Miller George, 1992)***](#_Miller,_G.A._(1922).) and *“What Are the Fundamental Concepts of Group Theory?”*[***(Burn Bob, 1996)***](#_Burn,_B._(1996).) *–* were introductory level texts to help me understand the core components of group theory. Whilst they were useful in this regard - helping me build my understanding - I found that once I started moving away from the idea of using groups, they quickly became redundant pieces of information. Additionally, because groups have been rarely used in classification models in the past (and after researching it has become evident to me why), it was difficult to find papers that addressed the specific area that I wanted them to cover. This meant that most of the papers or journals were full of information that either didn’t help me or didn’t apply to my topic.

Another issue I had was that group theory is a very young concept mathematically: especially in comparison to the information systems and theory papers I was also reading. This means that due to various technological advances (such as mainstream computer usage as well as the desire for artificial intelligence in the 80s and 90s), methods have changed a lot in the time between the two types of paper being produced. Group theory is a largely undiscovered field that has rapidly changing innovations from [***Mackey George, 1973***](#_Mackey,_G.W._(1973).)up to the present day. On the other hand, classification theory has only very recently been re-explored after years of uninterest (largely caused by the fact that the systems we developed long ago still work very effectively)

Overall, group theory was a very interesting topic area for me to explore and built up my knowledge (in set theory as well), even though it ultimately wasn’t the right fit for my project.

# ***Information systems***

Most of my early research time was placed into exploring pre-existing information systems to look at how they work and how effective they were. This area of research is essentially applied classification theory, so there is a large overlap between the theory and the implementations that I looked at. I studied some methods of data retrieval and how they worked - which was very helpful in guiding me in creating my own system. However, the largest portion of papers on this topic are written by bibliographers (with the specific intention of analysing books and library storage) so whilst there is a relation, I tended to get more conceptual ideas out of some of these applied papers (for example: [***Farradane Jason, 1950***](#_Farradane,_J.E.L._(1950).)). Most of the sources that I looked at were academic papers written by the same small collection of authors. A good portion of the content that I studied was following on from a previous work by the same author so there was a limited number of fresh perspectives that I could get on this topic. Nonetheless, I never felt as though the influx of ideas went stale and there were always new concepts for me to explore.

# ***Classification theory***

I found most of these papers after I had done a large portion of my initial research, when I was finally in a position to understand them. These papers were influential in the decisions that I made in my dissertation, and I often adapted ideas that I read about to make my model.

# ***Miscellaneous***

Inside this bracket I have collected a bunch of papers and other resources that helped me in developing ideas for my dissertation. Most of these only outlined formulae or methods and so couldn’t really be placed into any of the above categories. For example, I looked at “*Deriving Normal Equation of Linear Regression Model”* [***(Bunyamin Hendra)***](#_Bunyamin,_H._(n.d.).)to explore linear regression in a form that I could implement algorithmically and that I could use to evaluate the quality of my model. This source was very helpful to me because it provided explanations as well as the derivation of the formula so that I could understand the method in a large amount of detail. I also looked at *“A Study on Multiple Linear Regression Analysis*” [***(Uyanık & Güler ,2013)***](#_Uyanık,_G.K._and)to learn about how to extend the number of dimensions employed in linear regression so that I could more appropriately implement it in my dissertation. These sources as well as some other short website extracts that I read were vital in helping me decide how to proceed with the later sections of my project and the evaluation.

# **Dissertation**

Is it possible for all foods to be categorised into just being either a salad, sandwich, or a soup? In short, the debate surrounding the classification of food could be considered a captivating puzzle that sparks both fascination and contemplation. While the practicality and sensibility of such an all-encompassing model can be questioned, the fundamental premise holds true: any assortment of edible objects can indeed be grouped and ascribed arbitrary labels. This concept is deeply rooted in the nature of objects and the malleability of language: Just as a French "pomme" is unequivocally recognized as an English "apple", the core identity of an object transcends the labels we assign to it. The essence of our question hence hinges on the intrinsic human drive to categorize and make sense of the world, rather than any scientific merit. However, in this process we explore many key concepts and ideas that have real world value.

In the early stages of my research, I delved into the exploration of pre-existing information systems, looking at their inner workings and gauging their effectiveness. This study of applied classification theory revealed a significant convergence between theoretical constructs and real-world implementations. What this means is that even if we are not aware consciously of these groupings – they permeate through our language and our lives. My investigations explored methods of data retrieval, offering insights that would later inform the creation of my own classification system[[2]](#footnote-3). While much of the literature in this domain is produced by bibliographers with a specific focus on book analysis and library organization, I did manage to find several researchers who viewed the more conceptual aspects of these applied papers [***(Rosch Eleanor, 1973)***](#_Rosch,_E.H._(1973).)***.***

Our understanding as humans of the properties ingrained in linguistic description will always bias our perspective. This is an easily testable claim: if you think of an object and some truths about it, by simply changing up the wording of those truths you shouldn’t be too far from an unverifiable claim or even a wildly inaccurate statement. This is seemingly because the linguistic structure surrounding any object is inherently tied to the context of its use – a good example of this would be the “tin box argument” which can be read about here: [***Talk: Tin box - Wikipedia***](https://en.wikipedia.org/wiki/Talk:Tin_box). How can you differentiate linguistic labels when the nature of speech is so far separated from that level of categorical logic? Cans, bottles, boxes, universes are all examples of containers: but when we look at such an abstracted definition then our whole understanding of the nature of reality completely breaks down. This is either a testament to the fragility of our mathematical definitions or to the strength and importance of language to define and provide meaning to objects in Plato’s realm of forms.

Whilst I will provide a full breakdown of my method, a key takeaway from this paper should be that of repetition and refinement (I will explain the generic form as well as diving into my own explorations). This is a crucial point to understand moving forward – although my model has been setup in a particular way, with arbitrary parameters that I have chosen, the reader is welcome to explore with their own algorithms, variables, and context. This also leads quite nicely into another critical point of this paper: Perspective.

For any model analysis there are always 4 outcomes. The first is that the variables and methods you have chosen, fit the context of your question because of a well-defined and clear mathematical correlation between the two. Secondly, the model might fit by a fluke – there might exist some multi-staged connection or hidden variable that you had not considered. Thirdly, the model does not fit because you chose the wrong parameters or did not optimise your algorithms. And lastly, the model does not fit because you asked the wrong question (the context of your model did not appropriately reflect the phenomenon you were trying to show). It is important to understand that all of these are valid conclusions and all need to be examined in your evaluation.

Generally, the process follows some simple steps. This paper will guide the reader through these steps one by one, at the same time explaining and reviewing key ideas, research points and relevant papers.

***Prepare your variables.***

One of the biggest shortfalls of creating any categorical model is incomplete information. This generally occurs when you need to examine an object in a non-controlled environment. In my case, this happens because we want the model to be easy for any to evaluate with limited tools on hand: ideally, we would know everything about our entity, as well as have an advanced embedded device optimised to perform exact calculations and comparisons. Unfortunately, we do not have all this information or the resources necessary to collect the data. This means we need to balance accessibility and accuracy: this all comes from our variable and algorithm selection.

The first step for constructing a model is defining, describing, and preparing your variables from the physical entity[[3]](#footnote-4). In our case, we will be using two photographs of each object, one from a top-down view and one from a flat side-on view. All the information we need can be extracted from these two images and common knowledge. Again, it is important to think about the availability of data: we would much prefer a full 3d render of every object we want to compare but since this is most definitely out of the realm of plausibility for most people, two pictures will have to suffice.

An example of the types of information that we could extract from these pictures is as follows: *Colour, Colour variance, Structure, Structure variance*. Finally, we could use our knowledge of components (ingredients) to finalise the model. It is a good idea to make note of what properties we want to contribute to our model and how they could be collected (this aids not only your own bigger-picture view but also makes your logic completely transparent for other people). An example for our project could be:

***Colour***

For the colour variable, nothing special needs to be done to either image.

***Colour Variance***

To measure the colour variance of the image, we need to calculate the mean R, G and B pixel values for the whole image, we then compare each pixel’s values to the mean. Take the average of these (because there are 3 variables) which then gets added to 127 on a greyscale value chart. The pixel then gets assigned that colour.

A diagram of a burger

Description automatically generated

***Structure***

The image needs to be silhouetted. Another copy also needs to be grey scaled for comparison later.

***Structure variation***

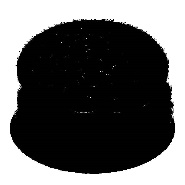
The image needs to be silhouetted and random points chosen on the perimeter such that lengths can be taken from the centre to them.

***Components***

A collage of a hamburger and spaghetti

Description automatically generatedThe reader should know what components make up the dish they are comparing. They do not need to know any measurements. If needed, a tool could be constructed to search the internet for the component ingredients.

A black silhouette of a cookie

Description automatically generated

***Classify***

Once all the data points for comparison have been selected and measured, we need to understand exactly what method we will be using for classification. We are going to be using a formation of “prototype” classification: this is a method in which classification is decided by comparison to a constant “archetypal example” of the class. All objects get compared in some way to this prototype and classified accordingly. We will be using prototype classification because it provides strong natural support for 2 major factors in our model: it is extremely easy to work with numerically (especially with our framework of fuzzy set theory) and it is quite easy to work with computationally [***(Rosch Eleanor, 1973)***](#_Rosch,_E.H._(1973).). Moreover, prototype classification is arguably the most intuitive to us as human beings, mimicking standard cognitive functions [***(Oliva et al, 2015)***](#_Vondrick,_C.,_Pirsiavash,).

For our similarity measure we will be using Tversky’s similarity index[[4]](#footnote-5). It neatly blends fuzzy set theory and prototype classifications into one comparison formula. The choice of this formula over any other is again an arbitrary decision but I have chosen to use due to the simplicity and elegance. Additionally, the formula has 2 tweakable parameters that control the “strength” of the prototype and variant: this means that the formula can be changed to be stricter towards what it accepts in a class – if you require less variation in each category.

Set theory is the branch of mathematics that explores collections and groupings of objects. Fuzzy set theory is an extension that allows each element a prescribed degree of membership to a set. Rather than just a binary choice in which items are either in or not in a collection – fuzzy sets allow for objects to be partially in a set (which we understand intuitively either as demonstrating some properties of the set but not all, or demonstration of a property to a lesser extent). Fuzzy set theory lends itself to classification very well because it allows for an easier framework to incorporate “grey areas” into classes: that is to say, groups can encompass a larger range of objects because they can specify that an object is only similar, to any given degree. It is immediately apparent how this can be useful for our project, food items such as lasagnes have key similarities to multiple classes (layers can mimic sandwiches, sauce consistency can mimic soups). There is much more freedom in this definition because it supports description of objects that do not neatly fit as a definite class. These sets are usually notated where represents our abstract object and represents its given membership value for the set [***(Roberts David, 1986)***](#_Roberts,_D.W._(1986).).

The full algorithm that we will be following has been provided at the end of the document, since it is not relevant to the core ideas of the research and is only for the interested reader or to provide clarifying context. Once a numeric value has been calculated from the procedure, we can then place it into the class structure in the form of a fuzzy set member . A key difference to note against Davids Robert’s work is that he expanded his initial paper to include what he called an “Anticommutative difference operator”, [***(Roberts David, 1987)***](#_Roberts,_D.W._(1987).) which we will be omitting for simplicity. Essentially Roberts wanted to be certain that class categorisation was unique and unambiguous, so he introduced the **“/”** operator for fuzzy sets which roughly translates into meaning “While not” in English. This meant that he could compare not only the similarity of set membership but also how separated this was from the other sets – thus showing how neatly fitting each object was.

A collage of many squares

Description automatically generatedPrototype theory in combination with our natural language view of objects leads to some very nice outcomes. Work done separately by both [***Aude Oliva***](#_Vondrick,_C.,_Pirsiavash,) and [***Eleanor Rosch***](#_Rosch,_E.H._(1973).) perfectly complement each other to paint a larger picture of the nature of objects that can be analysed computationally. Both Rosch’s primitive study and Oliva’s modern digital discovery indicate that “similar” entities (that is, entities from the same base/larger class) have many structural similarities. Oliva explored the idea that overlayed images from the same subclass seem to make an archetypal image, and Rosch showed that by taking many 2d representations of objects and drawing the overlap in them ends up with the ideal form.

A person standing in a forest

Description automatically generated

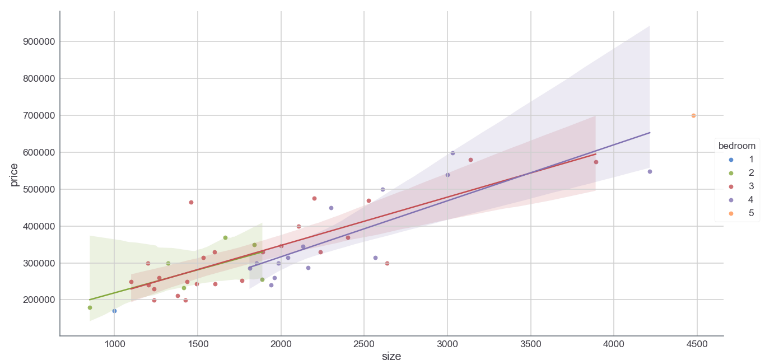
[“Average” image of people in Cambridge](https://groups.csail.mit.edu/vision/torralbalab/)

"Average" images, [Antonio Torralba (2004)](#_Vondrick,_C.,_Pirsiavash,)

***Analyse - Matrices***

Once we have constructed a model, we need to test its effectiveness: a process known as evaluating the model. One of the ways we will be doing this is by employing a technique commonly used in machine learning; however, we will be doing the process in reverse. Linear regression is a statistician's tool that constructs a model based on provided variables and their corresponding data, then using this to predict future results. We can keep the method largely the same although we are more interested in the model the algorithm produces than what it predicts. This is because we will be using the variable coefficients provided as a reference point for determining how fair, encompassing and – as a result – valid the classifications we formulated are. Two different methods for analysis will also be discussed.

Linear regression takes in a set of variables and the real world reflected results that we observe when these variables have certain values. A good example would be a house price model: we may take in variables about crime rate, number of nearby facilities and journey time to key locations and we look at how house prices change as these variables change. We use historic results (such as how expensive the house was yesterday) to help us predict future results (what it will cost in the future). Regression is used in many business sectors as well as in computer vision and machine learning and is a powerful method of analysing data [***(Uyanık & Güler, 2013)***.](#_Uyanık,_G.K._and)

A black text with white text

Description automatically generated with medium confidence

[Medium.com - Multivariate linear regression](https://miro.medium.com/v2/resize:fit:2000/format:webp/1*J6_stt6fkmVHkgsnUr0PhA.png)

As with the “classify” step, a full method has been provided below. Unlike the previous method, this is much more general and less tailored to our specific context which means it can be applied to a much wider class of problem. The method should be consistent with the method for regular multivariate linear regression for model prediction up until step 4, where we divert and evaluate differently.

***“Algorithm for model evaluation:***

1. *Place all the class similarity values in a n X m matrix where n represents the number of fields of data being included (These are the data points you are testing, e.g., burger, hotdog, ice cream etc) and m is equal to the number of classes these objects are being categorised into (e.g., salad, soup, sandwich etc)*
2. *Set the “y” matrix to the expected values, so in this case when you have x similarity values against each other you expect the model to place the object in the class with the highest similarity.*
3. *Perform standard multivariate linear regression:*
4. *Add a “bias” column to the matrix in the first position, filled with ones.*
5. *Use the normal equation to solve for coefficient values.*
6. ***At this stage, instead of using those coefficients to predict future data, we perform a comparison of them.*** *We can do this because theoretically if each class variable has an equal weight (since they all use the exact same process to calculate their values) then we would expect the coefficients to have equal weight (within reason, obviously a large amount of data would need to be given first). If there is a large discrepancy between one of the weights, then we know that the class it refers to holds less relevance than the others. It also means that the model itself isn’t properly balanced and the arbitrarily chosen classes are not the best fit together.”*

A math equations and numbers

Description automatically generated with medium confidence

In theory, if each class variable has an equal weight (since they all use the exact same process to calculate their values) then we would expect the coefficients to have equal weight given a large enough sample size. This isn’t surprising, especially for those with even a small mathematical background, as objects in nature often form according to a normal distribution. If there is a large discrepancy between one of the weights, then we know that the class it refers to holds less of a weight over the final model than the others. It also means that the model itself can’t be properly balanced and the arbitrarily chosen classes are not the best fit when taken together.

***Analyse – Graphs***

Another method we can use to verify the validity of our model is to use a serial axis plot (sometime called parallel coordinate plots) for multidimensional data. We can plot how many 1D property values compare to the same property in other entities – this helps us visualise how each entity compares to the others and where it sits in the larger model. To interpret these graphs, we use a continuous line of joined points to represent one entity: the colour of the line shows what class it fits into and the height of the line at each serial axis gives us an idea of the value of the given property.

A drawback to these graphs is that each line is not labelled, which means we can’t pick out any individual object. This is not a problem when being used to show the overall fit of the model but could be an issue if you wanted to take a more detailed view of any object. If we wanted to analyse these graphs to check the fit of our model, we would look for colour bunching – we want every line of a certain colour to fit closely with each other on each separate axis. This ensures objects with wildly different properties haven’t been placed into an inappropriate class. We then look for overlapping objects: where lines cross into groups of lines of a different colour[[5]](#footnote-7). This indicates that the property we are measuring on that axis is not entirely separated by the classes that we chose – naturally we want the property to be cut neatly at a certain point because this provides us with very helpful information about the borders of properties. For instance, it could give us a bound for what [***Gwen Malmquists***](#_Malmquist,_G._(2022).) “soupiness lemma” is.[[6]](#footnote-8)

Chart, radar chart

Description automatically generatedFinally, we need to check our graph for large gaps, this could suggest that our model is overfitting, and we need to include more class structures to cover the range of values that do not seem to have any food items fit into them. However, this could also be caused by a favouring in nature and linguistics to highlight certain attributes and cannot be relied on entirely to provide useful information about the system.

[Higher dimensional plotting - ggmulti](https://great-northern-diver.github.io/ggmulti/articles/highDim.html)

***Analyse – Analets***

An “analet” is a unit of information comprised from various atomic entities known as “isolates” and the operators that apply to them [***(Farradane Jason, 1950)***](#_Farradane,_J.E.L._(1950).). They were first constructed by Jason Farradane in 1950 and are used to organise knowledge. Since analets represent logical facts and category labels they are well suited to representing classification problems such as ours. Describing how to construct analets and the deeper implications of them is beyond the scope of this paper, although it is important to note that they are available as an option if you feel they would be appropriate for a project. Analets suit themselves to pattern matching problems and can be used to find and extract certain properties from a model – this is especially useful if you want to trace what part of a model is not working (for example, why all entities display a certain attribute, or that your similarity parameters are incorrect) and can be easily combined with using a parallel coordinate plot to make stunning model visualisations.

Analets are strings of isolates and operators and so fit into graph-based representations very nicely. We may want to show the analets in this form to make it easier to understand them as well as computationally because it enables us to perform certain graph theory algorithms (such as depth/breadth first search for comparing patterns).

A diagram of different types of equations

Description automatically generated with medium confidence

Graph representations of 3 example analets given by Farradane in "A scientific theory of classification". This are: “History of western philosophy”, “American owned factories in England” and “A comparison of economics in capitalism and socialism, respectively”.

A diagram of a triangle

Description automatically generated

An example of an analet graph for a very simplistic model of a ham sandwich. Much easier to understand than the corresponding string: **Sandwich/(butter/;sides(inside,both,[ham/+enclosed]/+closed/:bread/(white**

***The MSPACE method***

The MSPACE method is a checklist tool to walk the classification scientist through the process of model application on generalised categorisation problems. In our project, we will be following a systematic process: Model Construction, Similarity Measure Selection, Variable Preparation, Classification Approach, Critical Analysis & Evaluation, and Effective Presentation. This approach, which we'll refer to as 'MSPACE', is designed to ensure a structured and comprehensive analysis of our data, ultimately leading to more robust and interpretable results.

Now that we've introduced the 'MSPACE' framework, let's delve into what each letter represents:

**M – Model construction:** Make a note of the entities and classes that you want to categorise. You could consider visually mapping out all the entities that you want to classify.

**S - Similarity Measure Selection:** Choose the appropriate similarity measure. Make sure this integrates smoothly with whatever base system you are using. For prototype theory, I recommend Tversky’s comparison measure or the symmetric similarity measure. The Jaccard index is also good for this[[7]](#footnote-9).

**P - Variable Preparation:** Prepare your variables for analysis. It may be helpful if you note all these variables down as well as how you will collect the data. You may want to present this information at the end of your project so that other people can understand where the numbers come from.

**A - Classification Approach:** Decide what base system you are going to use and the other frameworks that integrate with this. My preference due to ease of intuition and overall simplicity would be either Prototype-variant theory with fuzzy sets or Exemplar theory.

**C - Critically Analyse and Evaluate:** Assess the model's performance and outcomes. You can do this both visually and numerically – regardless of the base system you choose I recommend checking weights using the matrix method. Remember to consider common points of failure in this section and account for the possibility of Gettier problems.

**E - Effective Presentation:** Finally, present your results with clarity and impact. Consider using some of the more visual representations discussed in the sections above such as serial coordinate axis plots or analet graphics.

Following this framework will provide an effective way of structuring your project as well as aid you in communicating your results. The MSPACE system has been developed to provide a standard way of presenting classification results, especially in environments in which key ideas need to be delivered to an audience outside of the mathematics field: It is intended to provide a general guideline to help people get into classification theory (with less intimidation) as well as set a precedent for people already familiar with the subject.

# **Evaluation**

The aim of this research paper was to present a comprehensive overview of the techniques used by scientists and mathematicians in the realm of classification theory. The MSPACE framework was introduced as a standard structure for classification projects, allowing scientists and mathematicians to work on a vast selection of diverse problems including those as far-fetched as *“salad, soup, sandwich”* classification. The generality of the framework does sacrifice some of the context specific aid, meaning that the framework cannot tell you what variables you need to pick – however it provides a loose guide to when you need to select these and the factors you need to consider.

Further research is needed to identify the effectiveness of the 3 analysis suggestions and their insightfulness. For example, it is unclear at this stage if analet analysis provides much merit beyond basic pattern recognition that could be conducted by a human, although the benefit of being computationally achievable shouldn’t be disregarded.

Ultimately, all classification problems contain an element of subjectivity: this is tied to the context of the problem, the linguistics of the problem and the nature of class/entity structures. As such, arguably the most vital part of any classification project is how well you can persuade the reader that your ideas are correct, and your assumptions are valid. Presentation is critical and hopefully this paper has shown how one can apply the MSPACE model to an example classification problem.

***APPENDIX A***

***Algorithm for object similarity:***

1. *Input 2 images “t”,”s” which are images of the item taken top down and from the side respectively*
2. *Apply pre-processing to the images: Resize them to fixed dimensions, centre them and remove and background.*
3. *Measure the colour similarity:*
4. *Compare raw pixel values between corresponding images from input and class S to a prescribed degree of accuracy and calculate percentage matching.*
5. *Create a colour distribution chart by calculating the mean colour and then the difference of each pixel from the mean. This gets converted into a range 0-127 then add/subtract this value from 127, which corresponds to a greyscale value (0-254)*
6. *Compare the percentage matching to a prescribed degree of closeness between the colour distribution chart for input and the corresponding image for class S*
7. *Calculate overall mean similarity by the equation: where: are the % raw pixel similarity for image and the % distribution similarity for image respectively.*
8. *Measure the structural similarity:*
9. *Calculate the % overlap of the silhouettes of the input images on the corresponding images for class S respectively.*
10. *Calculate the mean length from the midpoint of the image some random points on the perimeter then divide this value by the range of the lengths you have got. The similarity between value for your input and for the corresponding image in class S is given by in which refers to the input image value and refers to the class image value.*
11. *The overall mean similarity is calculated by the equation: where: are the % overlap similarity for image and the % mean length similarity for image respectively.*
12. *Measure the ingredient similarity:*
13. *A list of ingredients is supplied or requested from the internet and the similarity is calculated by the number of shared ingredients/sum of total unique ingredients*
14. *The overall similarity is calculated by the equation: where: is the % ingredient similarity for image*
15. *The total similarity for object O to class S is: where: are the % colour similarity for object , the % structural similarity for object and the % ingredient similarity for object against class S respectively.*
16. *Repeat this process for all classes.*

# **Bibliography**

# Bencivenga, E. (1976). Set Theory and Free Logic. *Journal of Philosophical Logic*, [online] 5(1), pp.1–15. Available at: <https://www.jstor.org/stable/30226131> [Accessed 15 Nov. 2022].

# Berlin, I. and Hardy, H. (1996). *Concepts and Categories: Philosophical Essays*. REV - Revised, 2 ed. [online] *JSTOR*. Princeton University Press. Available at: [https://www.jstor.org/stable/j.ctt46n3tc](https://www.jstor.org/stable/j.ctt46n3tc%20) [Accessed 27 Feb. 2023].

# Borel, A. and Harish-Chandra (1962). Arithmetic Subgroups of Algebraic Groups. *Annals of Mathematics*, [online] 75(3), pp.485–535. Doi: <https://doi.org/10.2307/1970210>.

# Bunyamin, H. (n.d.). *Deriving Normal Equation of Linear Regression Model*. [online] hbunyamin.github.io. Available at: <https://hbunyamin.github.io/machine-learning/Normal_equation/> [Accessed 25 Feb. 2023].

# Burn, B. (1996). What Are the Fundamental Concepts of Group Theory? *Educational Studies in Mathematics*, [online] 31(4), pp.371–377. Available at: [https://www.jstor.org/stable/3482970](https://www.jstor.org/stable/3482970%20) [Accessed 17 Nov. 2022].

# Culbert, S.S., Rosch, E. and Lloyd, B.B. (1980). Cognition and Categorization. *The Modern Language Journal*, [online] 64(2), p.284. Doi: <https://doi.org/10.2307/325356>.

# Farradane, J.E.L. (1950). A SCIENTIFIC THEORY OF CLASSIFICATION AND INDEXING AND ITS PRACTICAL APPLICATIONS. *Journal of Documentation*, 6(2), pp.83–99. Doi: <https://doi.org/10.1108/eb026155>.

# Farradane, J.E.L. (1952). A SCIENTIFIC THEORY OF CLASSIFICATION AND INDEXING: FURTHER CONSIDERATIONS. *Journal of Documentation*, 8(2), pp.73–92. Doi: <https://doi.org/10.1108/eb026182>.

# Gottwald, S. (2006). Universes of Fuzzy Sets and Axiomatizations of Fuzzy Set Theory. Part II: Category Theoretic Approaches. *Studia Logica: An International Journal for Symbolic Logic*, [online] 84(1), pp.23–50. Available at: [https://www.jstor.org/stable/20016819](https://www.jstor.org/stable/20016819%20) [Accessed 15 Nov. 2022].

# Hamlyn, D.W. (1959). Categories, Formal Concepts and Metaphysics. *Philosophy*, [online] 34(129), pp.111–124. Available at: [https://www.jstor.org/stable/3748729](https://www.jstor.org/stable/3748729%20) [Accessed 14 Nov. 2022].

# Jimenez, S., Becerra, C., Gelbukh Cic-Ipn, A., Dios Bátiz, A. and Mendizábal (2013). SOFTCARDINALITY-CORE: Improving Text Overlap with Distributional Measures for Semantic Textual Similarity. [online] 1, pp.194–201. Available at: <https://aclanthology.org/S13-1028.pdf> [Accessed 14 Mar. 2023].

# Lee, H.N. (1931). THE MEANING OF THE NOTATION OF MATHEMATICS AND LOGIC. *The Monist*, [online] 41(4), pp.594–617. Available at: [https://www.jstor.org/stable/27901326](https://www.jstor.org/stable/27901326%20) [Accessed 22 Nov. 2022].

# Löwe, B. and Welch, P.D. (2001). Set-Theoretic Absoluteness and the Revision Theory of Truth. *Studia Logica: An International Journal for Symbolic Logic*, [online] 68(1), pp.21–41. Available at: [https://www.jstor.org/stable/20016296](https://www.jstor.org/stable/20016296%20) [Accessed 8 Nov. 2022].

# Mackey, G.W. (1973). Group Theory and Its Significance for Mathematics and Physics. *Proceedings of the American Philosophical Society*, [online] 117(5), pp.374–380. Available at: [https://www.jstor.org/stable/986606](https://www.jstor.org/stable/986606%20) [Accessed 17 Nov. 2022].

# Malmquist, G. (2022). *Salad Theory*. [online] <https://saladtheory.github.io>. [Accessed 10 Nov. 2022].

# Miller, G.A. (1922). Easy Group Theory. *The Scientific Monthly*, [online] 15(6), pp.512–519. Available at: [https://www.jstor.org/stable/6660](https://www.jstor.org/stable/6660%20) [Accessed 17 Nov. 2022].

# Parsons, J. (1996). An Information Model Based on Classification Theory. *Management Science*, [online] 42(10), pp.1437–1453. Available at: [https://www.jstor.org/stable/2634376](https://www.jstor.org/stable/2634376%20) [Accessed 8 Nov. 2022].

# Roberts, D.W. (1986). Ordination on the Basis of Fuzzy Set Theory. *Vegetatio*, [online] 66(3), pp.123–131. Available at: [https://www.jstor.org/stable/20037322](https://www.jstor.org/stable/20037322%20) [Accessed 15 Nov. 2022].

# Roberts, D.W. (1987). An anticommutative difference operator for fuzzy sets and relations. *Fuzzy Sets and Systems*, 21(1), pp.35–42. Doi: <https://doi.org/10.1016/0165-0114(87)90150-3>.

# Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M. and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), pp.382–439. Doi: <https://doi.org/10.1016/0010-0285(76)90013-x>.

# Rosch, E.H. (1973). Natural categories. *Cognitive Psychology*, 4(3), pp.328–350. Doi: <https://doi.org/10.1016/0010-0285(73)90017-0>.

# Ross, B.H. and Spalding, T.L. (1994). Concepts and Categories. *Thinking and Problem Solving*, pp.119–148. Doi: <https://doi.org/10.1016/b978-0-08-057299-4.50010-4>.

# Russel, B. (1908). Mathematical Logic as Based on the Theory of Types. *American Journal of Mathematics*, [online] 30(3), pp.222–262. Doi: <https://www.jstor.org/stable/2369948>.

# Stanley, J. (2008). Knowledge and Certainty. *Philosophical Issues*, [online] 18, pp.35–57. Available at: [https://www.jstor.org/stable/27749898](https://www.jstor.org/stable/27749898%20) [Accessed 14 Nov. 2022].

# Stein, N. (2016). Causes and Categories. *Noûs*, [online] 50(3), pp.465–489. Available at: [https://www.jstor.org/stable/26631401](https://www.jstor.org/stable/26631401%20) [Accessed 14 Nov. 2022].

# Uyanık, G.K. and Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, 106(1), pp.234–240. Doi: <https://doi.org/10.1016/j.sbspro.2013.12.027>.

# Vickery, B.C. (1953). SYSTEMATIC SUBJECT INDEXING. *Journal of Documentation*, 9(1), pp.48–57. Doi: <https://doi.org/10.1108/eb026190>.

# Vickery, B.C. (1955). DEVELOPMENTS IN SUBJECT INDEXING. *Journal of Documentation*, 11(1), pp.1–11. Doi: https://doi.org/10.1108/eb026209.

# Vickery, B.C. (1963). SCIENTIFIC INFORMATION: PROBLEMS AND PROSPECTS. *Minerva*, [online] 2(1), pp.21–38. Available at: [https://www.jstor.org/stable/41821596](https://www.jstor.org/stable/41821596%20) [Accessed 29 Nov. 2022].

# Vondrick, C., Pirsiavash, H., Oliva, A. and Torralba, A. (2015). *Learning visual biases from human imagination*. [online] Available at: [https://proceedings.neurips.cc/paper/2015/file/8f53295a73878494e9bc8dd6c3c7104f-Paper.pdf](https://proceedings.neurips.cc/paper/2015/file/8f53295a73878494e9bc8dd6c3c7104f-Paper.pdf%20) [Accessed 2 Mar. 2023].

# Zagzebski, L. (1994). The Inescapability of Gettier Problems. *The Philosophical Quarterly*, 44(174), p.65. Doi: <https://doi.org/10.2307/2220147>.

1. Note that throughout this paper, the author uses the words item, entity, element and object interchangeably although strictly speaking, they should be split into two: entities, objects, articles, etc have properties (as well as relations - [***Farradane Jason, 1950***](#_Farradane,_J.E.L._(1950).)), Whereas items and elements are objects that belong to collections. [↑](#footnote-ref-2)
2. Technically, it would be inaccurate to say that I devised a system. It would be better understood as building the core principles that can be applied to create such a system. [↑](#footnote-ref-3)
3. On a technical note, the first step is to decide what type of model you want to use (this is usually not given much thought as you would pick one that is easiest to work with given the context of your problem). We will be using a prototype-based model which will be discussed in more detail later. [↑](#footnote-ref-4)
4. We are going to be using a symmetric variant of the formula to satisfy the requirement of similarity measure: [Tversky index - Wikipedia](https://en.wikipedia.org/wiki/Tversky_index) [↑](#footnote-ref-5)
5. Although if lines of every colour group near a certain point, then it might just represent a universal truth about the property you are measuring which is not necessarily a bad thing. [↑](#footnote-ref-7)
6. The hypothetical differentiation point of solids into liquids. This is a value between 0 and 1 that dictates how “soupy” a food is. [↑](#footnote-ref-8)
7. [Jaccard index - Wikipedia](https://en.wikipedia.org/wiki/Jaccard_index) [↑](#footnote-ref-9)